Journal of Statistical Planning and Inference 179 (2016) 36–59

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A regularized profile likelihood approach to covariance matrix estimation

Samprit Banerjee [a], Stefano Monni [b,*], Martin T. Wells [a]

[a] Cornell University, United States
[b] American University of Beirut, Lebanon

ARTICLE INFO

Article history:
Received 23 May 2015
Received in revised form 10 June 2016
Accepted 14 June 2016
Available online 28 June 2016

Keywords:
Covariance matrix
Eigenvalues
Isotonic regression
Laplace approximation
Modified profile likelihood

ABSTRACT

Two new orthogonally equivariant estimators of the covariance matrix are proposed. The estimates of the population eigenvalues are isotonized maximum likelihood estimates of the modified profile likelihood obtained from the Wishart distribution, in one case, and of a penalized form of such a likelihood function, in the other, with a penalty that constrains the trace of the sample covariance matrix. Properties of these estimators are studied and numerical risk comparisons with six other well-known estimators are presented to demonstrate the robustness of the proposed estimators for various real and simulated covariance structures.

1. Introduction

We observe N independently distributed p-dimensional random vectors x_i^T, $i = 1, \ldots, N$, drawn from a multivariate normal distribution with zero mean and unknown covariance matrix Σ, with $N > p$. This paper is concerned with the problem of estimating Σ from the data $X = (x_1, \ldots, x_N)^T$ (an $N \times p$ matrix). The usual unbiased estimator of the covariance matrix is the sample covariance matrix $S = \sum_i^N (x_i - \bar{x})(x_i - \bar{x})^T/(N-1)$, with $\bar{x}$ being the p-vector of the sample mean. Let $S = HLH^T$ and $\Sigma = O\Lambda O^T$ be the spectral decompositions of S and Σ, so that H and O are the matrices of an orthonormal basis of eigenvectors (elements of the real orthogonal group $O(p)$) and $L = \text{diag}(l)$ and $\Lambda = \text{diag}(\lambda)$ are the diagonal matrices of the eigenvalues $l = (l_1, \ldots, l_p)$ and $\lambda = (\lambda_1, \ldots, \lambda_p)$, respectively. Throughout this paper we will assume that the sample and population eigenvalues are ordered, *i.e.* $l_1 > l_2 > \cdots > l_p > 0$ and $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$.

The sample covariance matrix, S, follows a p-dimensional Wishart distribution with $n = N - 1$ degrees of freedom and expectation Σ. One disadvantage of S is that its eigenspectrum is more spread out than the population spectrum unless $\gamma = p/N$ is negligible, *i.e.* the largest eigenvalue of S has a positive (upward) bias and the smallest has a negative (downward) bias. This has long been known and can be proven in a number of ways. In a language that will often be employed in this paper we can say that it is a direct consequence of the fact that the vector $E(l) = (E(l_1), \ldots, E(l_p))$ of the expectations of the eigenvalues of S majorizes the vector λ of the eigenvalues of Σ. In other words, $\sum_{i=1}^J E(l_i) \geq \sum_{i=1}^J \lambda_i$ for any $J = 1, \ldots, p-1$, and $\sum_{i=1}^p E(l_i) = \sum_{i=1}^p \lambda_i$. This is a corollary of a more general theorem for random Hermitian matrices due to Cacoullos and Olkin (1965) (see also Marshall et al. (2011)).

Many studies have been dedicated to finding improved covariance matrix estimators that are less affected by the sample eigenvalue dispersion issue or that are better than S in terms of risk. One approach looks for estimators that are equivariant

* Corresponding author.
E-mail addresses: sab2028@med.cornell.edu (S. Banerjee), sm150@aub.edu.lb (S. Monni), mtw1@cornell.edu (M.T. Wells).

http://dx.doi.org/10.1016/j.jspi.2016.06.004
0378-3758/

under a group of transformations that acts on the sample space. Generally, the adequacy of the resulting estimator is studied with respect to loss functions that are invariant under the same group so that the entire decision problem is indeed invariant. Another approach imposes constraints on the sample covariance matrix in the form of regularization (for a review see Pourahmadi, 2011). In this article, we follow the first approach and consider estimators that are functions of S and are equivariant under the orthogonal group $O(p)$. Namely, when the sample space is acted on by the orthogonal group according to $X^T \mapsto OX^T$, so that $S \mapsto OSO^T$, $\Sigma \mapsto O\Sigma O^T$, we require the estimators to transform as $\phi(S) \mapsto \phi(OSO^T) = O\phi(S)O^T$. Orthogonally equivariant estimators $\phi(S)$ are of the form $\hat{\Sigma} = \phi(HLH^T) = H\hat{\Lambda}H^T$, with $\hat{\Lambda} = \phi(L) = \text{diag}(\hat{\lambda})$ the matrix of the estimators $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ of the population eigenvalues $\lambda = (\lambda_1, \ldots, \lambda_p)$. Hence, equivariant estimators differ from each other only in how they estimate the population eigenvalues. Geometrically, the level set $Q_{\hat{\Sigma}}(z) = \{z : z^T\hat{\Sigma}z = c, c > 0, z \in \mathbb{R}^p\}$, of the positive definite quadratic form of such an estimator $\hat{\Sigma}$ is an ellipsoid with the same principal axes as the level set of the quadratic form of S, but different lengths. Thus the level sets of the orthogonally invariant estimators are a reshaping of the sample covariance matrix ellipsoids but have the same orientation. One of the best (in terms of risk) orthogonally equivariant estimators was proposed by Stein (1975, 1977), who obtained $\hat{\lambda}$ by (approximately) minimizing the unbiased estimator of the risk under Stein's loss (denoted by LS) (Stein, 1956) (see Section 5 for definition). Problems with such estimator were remedied by adjusting the eigenvalue estimates so as to make them positive, and, via isotonic regression, ordered. We will henceforth refer to the resulting estimator as Stein's estimator. A similar estimator was also proposed by Haff (1991), where $\hat{\lambda}$ are obtained by minimizing the Bayes risk under LS (and a second one under the quadratic loss, denoted by LQ, see Section 5). Other orthogonally equivariant estimators proposed in the literature include Dey and Srinivasan (1985), Perron (1990, 1992), Takemura (1984) (see also the review by Pal, 1993), and more recently Ledoit and Wolf (2004, 2012, 2014, 2015). In Ledoit and Wolf (2012) the authors introduce a non-linear shrinkage estimator, where eigenvalues are obtained by applying a non-linear transformation to the sample eigenvalues. They show how such estimator provides improvements over S and performs better than the linear shrinkage estimator (Ledoit and Wolf, 2004) constructed by shrinking all sample eigenvalues with the same intensity toward their mean. The nonlinear shrinkage approach was further extended to the $p > N$ case in Ledoit and Wolf (2015, 2014) where estimators were proposed that are asymptotically optimal in the class of orthogonally equivariant estimators under the Frobenius loss (see Section 5) and under LS, respectively.

Naturally, groups of transformations other than the orthogonal group can be considered. For general linear transformation (invertible matrices), the equivariant estimators are known to be proportional to S (Stein, 1956). Much consideration has also been given to transformations corresponding to lower triangular matrices with positive diagonal elements. The corresponding group $\mathcal{G}_T^+$ acts transitively on the space $\mathcal{S}_p^+$ of positive definite symmetric matrices with the implication that the risk of a $\mathcal{G}_T^+$-equivariant estimator under a $\mathcal{G}_T^+$-invariant loss is constant. For example, James and Stein (1961) and Stein (1956) obtained one such estimator, which has since been shown to be dominated (under Stein's loss) by the orthogonally equivariant Stein's estimator, among others. In addition, orthogonally equivariant estimators are generally more interesting because, contrary to the $\mathcal{G}_T^+$-equivariant ones, they do not depend on the coordinate system in the sample space, as a change of basis is achieved by elements of $O(p)$.

Symmetry is not the only principle that can be employed to constrain a problem. Among the more recent literature, banding or tapering the sample covariance matrix is used by Bickel and Levina (2008) to define regularized estimators. Regularization via banding or tapering assumes the variables have a natural ordering and thus it is appropriate only for problems with a precise underlying structure. Bickel and Levina also employ hard thresholding to regularize the sample covariance matrix (Bickel and Levina, 2008) and obtain an estimator which is equivariant under permutations (which form a subgroup of the orthogonal group) and thus does not require the data to be ordered. They show that both the banded and thresholded estimators are consistent in the operator norm as long as $(\log p)/N \to 0$. In this paper, however, we will focus on the finite sample properties of covariance matrix estimation, particularly in the high-dimensional setting, when N and p are both large but $p < N$.

The starting point of our construction is the joint distribution of the sample eigenvalues which is obtained by integrating out the eigenvectors with respect to the Haar measure of $O(p)$ (Muirhead, 2009). In our setting, the parameter space, which is $\mathcal{S}_p^+$, coincides with the sample space, if we look at the data via their sufficient statistic S. The Wishart density $p(S|\Lambda, O)$ (with respect to the invariant measure $\mu(dS) = dS/|S|^{(p+1)/2}$) is such that $p(G^{-1}(S)|\Lambda, O) = p(S|\Lambda, GO)$ with $G \in O(p)$, where the $O(p)$-action is defined as $G(A) = GAG^T$, $A \in \mathcal{S}_p^+$. It follows that the marginal density of the sample eigenvalues depends only on the population eigenvalues. The marginal likelihood $f(\Lambda|L)$ of the invariant parameter Λ can also be written as an integral over the Haar measure of the parameter O, $\int_{O(p)} f(\Lambda, O|S)(dO)$. The complete form of this marginal likelihood is however intractable. For this reason, we consider its Laplace approximation, given in Eq. (2). This approximate likelihood is the modified profile likelihood (Barndorff-Nielsen, 1983; Muirhead, 2009; Rajaratnam, 2006). As eigenvalue estimators, we propose the isotonic regression of the n^{-1}-order maximizers of this profile likelihood. Isotonic regression is considered to guarantee the correct ordering. We show that these estimates are positive, so there is no need for an additional computational step to obtain positivity, as is necessary in the implementation of Stein's estimator. We then propose a penalized estimator, which is the isotonic regression of the maximizers of a penalized likelihood function, where the penalty can constrain the trace of the covariance matrix estimator to be a fraction of the trace of the sample covariance matrix. We investigate finite sample properties of this estimation procedure through risk comparisons with other